Published online 29 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D1251–D1260***
doi:10.1093/nar/gks1279

NCBI Bookshelf: books and documents in life sciences and health care

Marilu A. Hoeppner*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 45 Center Drive, Bethesda, MD 20892, USA

Received October 7, 2012; Revised and Accepted November 8, 2012

ABSTRACT

Bookshelf (http://www.ncbi.nlm.nih.gov/books/) is a full-text electronic literature resource of books and documents in life sciences and health care at the National Center for Biotechnology Information (NCBI). Created in 1999 with a single book as an encyclopedic reference for resources such as PubMed and GenBank, it has grown to its current size of >1300 titles. Unlike other NCBI databases, such as GenBank and Gene, which have a strict data structure, books come in all forms; they are diverse in publication types, formats, sizes and authoring models. The Bookshelf data format is XML tagged in the NCBI Book DTD (Document Type Definition), modeled after the National Library of Medicine journal article DTDs. The book DTD has been used for systematically tagging the diverse data formats of books, a move that has set the foundation for the growth of this resource. Books at NCBI followed the route of journal articles in the PubMed Central project, using the PubMed Central architectural framework, workflows and processes. Through integration with other NCBI molecular databases, books at NCBI can be used to provide reference information for biological data and facilitate its discovery. This article describes Bookshelf at NCBI: its growth, data handling and retrieval and integration with molecular databases.

INTRODUCTION

Bookshelf (http://www.ncbi.nlm.nih.gov/books/) is an online service providing free access to the full text of books and documents in life sciences and health care, built and maintained by the National Center for Biotechnology Information (NCBI) within the National Library of Medicine (NLM). Bookshelf complements the other biomedical literature resources at NCBI: PubMed, the database of citations and abstracts, PubMed Central (PMC), which provides free access to the full text of journal articles and author manuscripts in biomedicine, and the NLM Catalog, which holds catalog information for books and journals as well as other materials such as audiovisuals. This article discusses the Bookshelf literature resource at NCBI, including factors that have contributed to its growth, how the data are structured, processed, stored and retrieved and its significance in the context of molecular information.

The core objectives for Bookshelf are (i) to further advance science and improve health care through the collection, exchange and dissemination of books and related documents in life sciences and health care; (ii) to provide a permanent stable archive for the collection and (iii) to provide free access to full text of this content. As a full-text literature hub in a network of genomic and molecular genetics resources at NCBI, Bookshelf has an additional goal to enable literature annotations for factual information found in the genomic and molecular databases at NCBI and to facilitate the discoverability of this information.

Bookshelf titles are of varied publication types, topics and authoring models, and include public domain works as well as works the copyright holders of which have granted NCBI permission for distribution. Bookshelf includes textbooks, monographs, health reports, documentation, website content and databases. Content spans the range of basic biology to advanced molecular biology, health care evidence reports and clinical guidelines to health care policy analysis and reports from high-throughput screening of small molecules for identification of disease-combating drugs to guidance for drug development. Content is selected for the collection based on the scope of the subject matter as defined by NLM's Collection Development Manual (http://www.nlm.nih.gov/tsd/acquisitions/cdm/), as well as on scientific and editorial quality criteria and the technical quality of the submitted files.

Bookshelf serves as an educational resource. It is a knowledgebase of information for undergraduate and

*To whom correspondence should be addressed. Tel: +1 301 496 4911; Fax: +1 301 480 2484; Email: hoeppner@ncbi.nlm.nih.gov

graduate students, scientists, health care professionals, patients and users seeking biomedical information. The free availability of Bookshelf content enables global access to this knowledgebase by users who might otherwise not have access to this data. The content is made freely available to readers by authors, editors and publishers who agree to participate in the project. The collection also includes reports and documents from US and international government agencies as well as organizations in the health sector. Publishers and content providers also benefit by wide distribution of their content to the general public, health care professionals and a population of students who will become the next generation of biomedical researchers, clinicians and teachers. Where participatory agreements permit, content from the collection is made available in the Open Access Subset (http://www.ncbi.nlm.nih.gov/books/about/openaccess/), in which XML, image and supplementary files are shared, allowing for redistribution and reuse of the content. Information for authors and publishers on selection criteria and the application process can be found at: http://www.ncbi.nlm.nih.gov/books/about/publishers/

Queries can be addressed to Bookshelf via the 'Write to Help Desk' link on every Bookshelf web page or via e-mail to: bookshelf@ncbi.nlm.nih.gov.

GROWTH OF BOOKSHELF

Bookshelf started in 1999 with the third edition of Molecular Biology of the Cell, Alberts *et al.* (1). Since then, Bookshelf has grown to its current count of 1373 titles as of August 2012 (Figure 1). The Bookshelf collection is relatively small compared with PubMed and PMC as determined either by the number of accession identifiers (IDs; accessible units of content) or by the absolute size for the different data types (see Table 1). To organize the diversity of publication types in Bookshelf, each title is assigned a Bookshelf resource type and one or more subjects. The list of resource types is shown in Table 2.

Three developments in Bookshelf have had a major impact on growth of the collection: (i) the development of the NCBI Book DTD (Document Type Definition) based on the DTDs of the Journal Article Tag Suite (JATS) and migration of XML data to this DTD (see 'Data', 'Format

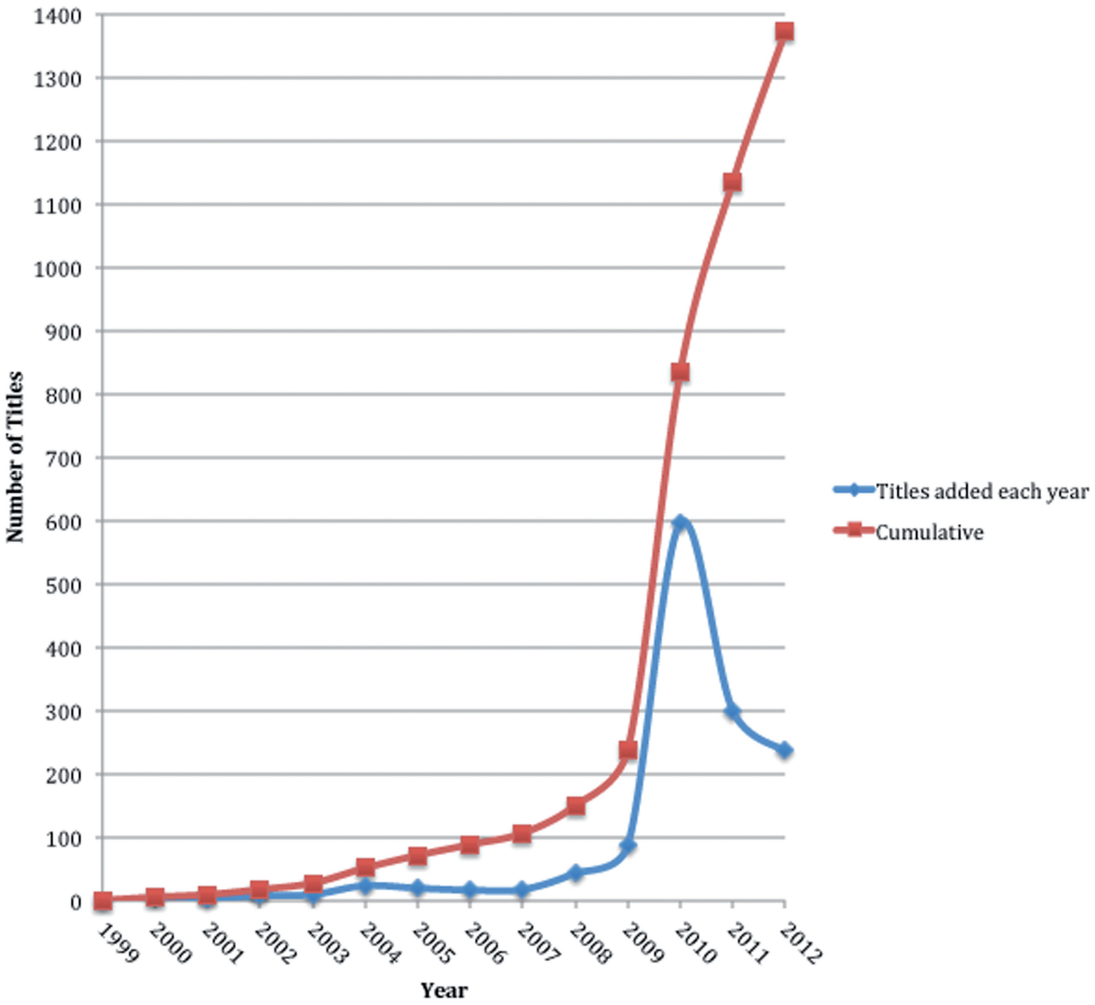


Figure 1. Growth of Bookshelf. The term 'title' represents a book in the database. The spike in the book title count in 2010 was due to data restructuring of the Health Services/Technology Assessment Texts (HSTAT; http://www.ncbi.nlm.nih.gov/books/NBK16710/) database content. Previously, reports in this database were not counted as individual titles.

Table 1. Size of books data compared with PMC and PubMed

Resource	No. of organizational unit	No. of accession IDs	XML	PDF	Scanned image	Image[a]	Media	Supplementary files
Bookshelf	1373 books	37 275	4 GB	6 GB	N/A	5 GB	3 GB	1 GB
PMC	~1090 Journals (full participation) ~230 Journals (NIH portfolio) ~1750 Journals (selective deposit)	~2 400 000	85 GB	2500 GB	970 GB	470 GB	210 GB	940 GB
PubMed	~25 000 Journals	~22 000 000	110 GB	N/A	N/A	N/A	N/A	N/A

Data as of 31 August 2012. Size data in gigabytes (GB) is approximate.
[a]Excluding high-resolution image files in tileshop database.

Table 2. Resource types in bookshelf

Type[a]	Description	Percentage of total
Report	Formal publication of scientific meeting or workshop, research investigation or area of study	83.8
Book	Text book or monograph	10.3
Collection	A title that holds references to a collection of related books or series	1.8
Database	A title containing content-related documents that are highly structured, analogous to database fields	1.8
Documentation	A title that describes a companion web or software application	2.3

[a]Each title in Bookshelf receives a type designation that describes the source material.

and structure' section), (ii) rendering books searchable in PubMed and (iii) the redesign of Bookshelf and integration into the NCBI portal application, which facilitates sharing of data across resources at NCBI (2,3).

DATA

Format and structure

The DTDs of the JATS were initially developed for tagging journal articles in the PMC project. JATS DTDs aim to capture the semantics of the content independent of the form in which it is delivered. Widely used in PMC and worldwide to tag journal articles, they are now a National Information Standards Organization (NISO) standard, NISO z39.96-2012 [(4) ANSI/NISO Z39.96-2012 JATS: Journal Article Tag Suite: Project Overview. Available at: http://www.niso.org/apps/group_public/project/details.php?project_id=93. Accessed 10 September 2012). Books and journal articles share many common features: metadata such as title, contributors, publisher information and publication dates, textual content with special characters and emphasis, images, tables, equations and reference lists. Early in the project, Bookshelf used a DTD based on the ISO 12083 (International Organization for Standardization) article DTD for tagging data. The tag set had to be continually modified as the project expanded, adding challenges to data management and rendering. To overcome these problems, the NCBI Book DTD (http://dtd.nlm.nih.gov/book/) was developed specifically for the Bookshelf project for tagging book content, and Bookshelf data was migrated into this DTD. It is modeled along the same design principles as the JATS DTDs, and uses many of same modules. Bookshelf XML data are currently tagged in the NCBI Book DTD, v2.3. The similarities between books and journal articles, and between their shared tag sets, have permitted Bookshelf to leverage the robust PMC architectural framework as well as existing PMC workflows and tools for handling the data. The NCBI Book DTD in the context of JATS has been discussed in detail (5).

Submission, XML conversion and storage

Tagging content semantically in XML is one of the most complex and costly operations for Bookshelf. To enable continued maintenance of the corpus of book data, and continued growth of Bookshelf, it has been necessary to balance the needs of the publisher with the resources of the Bookshelf by streamlining the number of submission formats. To this end, Bookshelf recently moved toward a requirement for data submission in semantically tagged XML, which permits partial or complete automation of data processing. XML data are submitted either in the NCBI Book DTD or in an alternate DTD (e.g. DocBook). When submission uses an alternate DTD, Bookshelf uses XSLT converters to transform the XML to the NCBI Book DTD format. For submission of data in NCBI Book DTD XML, tagging guidelines (http://www.ncbi.nlm.nih.gov/pmc/pmcdoc/tagging-guidelines/book/style.html) have been developed and are based on similar tagging guidelines for JATS DTDs. These guidelines are intended to guide proper tagging practice through tagged samples, to reduce the variability in tagging data elements and facilitate data exchange. A subset of Bookshelf projects that require frequent updates are authored in a specialized Microsoft Word template, which uses styles to semantically tag the document elements, such as the title and author list. The documents are converted to XML using the in-house NCBI Word Converter tool, which uses the eXtyles product (Inera, Inc.) for reference processing. Documents are updated in Microsoft Word and reprocessed using the Word Converter. Legacy projects involving print publications

are submitted in PDF format and are converted by third-party vendors to NCBI Book DTD XML. FTP is the main portal for data submission.

XML, image, PDF and supplementary files are stored in the Bookshelf content management system (CMS), built in-house for Bookshelf. In the CMS, book data are checked for validation against the DTD, conformance to an in-house style checker (which runs additional checks beyond XML validation to ensure data quality) and to ensure that all files associated with the book are available (see Quality Checks mentioned later in the text). The CMS is a destination hub for NCBI Book DTD XML data that is received through a number of workflows and the staging area for ingest and subsequent processing of book data. All XML content stored in the CMS is in the form of a single XML document with the root element <book>.

Processing

The main steps of data processing are (i) ingest, (ii) 'chop-it-up' process, (iii) text and image processing and (iv) PDF build (see Figure 2). Ingest begins with downloading XML, image and supplementary files from the CMS onto the file system and then bundling them to create a tar file. Chop-it-up and text processing involve XSLT transformation on XML data, creating XML output. During the chop-it-up process, the single independently validating NCBI Book DTD XML document with root element <book> is chopped up into independently validating XML documents with root element <book-part>; i.e. the book is divided into stand-alone book units such as front-matter sections, chapters, appendices or reference lists. Book metadata is carried into every book part, and cross-references to book parts are resolved in a way that the files can validate. The creation of article-like <book-part> XML files from the <book> XML has provided the basis for using the PMC workflows and tools for Bookshelf data processing. Text processing and image conversion occur in parallel. For text conversion, the software resolves named entities, handles special or custom characters and custom math, validates XML and runs the style checker. For image conversion, the software that runs on open-source ImageMagick (ImageMagick Studio) determines image dimensions and properties such as size, type and resolution, resizes images per Bookshelf specifications and creates for each image a thumbnail, a web-resolution JPEG file and a high-resolution JPEG file (if the source files were of high resolution). PDFs are created for book chapters if not provided by the content provider and if their creation and display in Bookshelf is permitted. The PDF build software uses the XML output of text conversion and creates a formatting object file, gathers image heuristics and resizes images so they are compatible with print layout. The Antenna House formatter (Antenna House, Inc.) creates the PDF from the formatting object file.

Loading to the database

The loading software identifies the XML files for addition or replacement and loads them to the database. It validates the data, performs checks for file types and associated files and resolves loading of files associated with each XML file, such as images, equations, multimedia and supplementary files. It parses the XML for key metadata information, such as book-part identifiers for storage in the

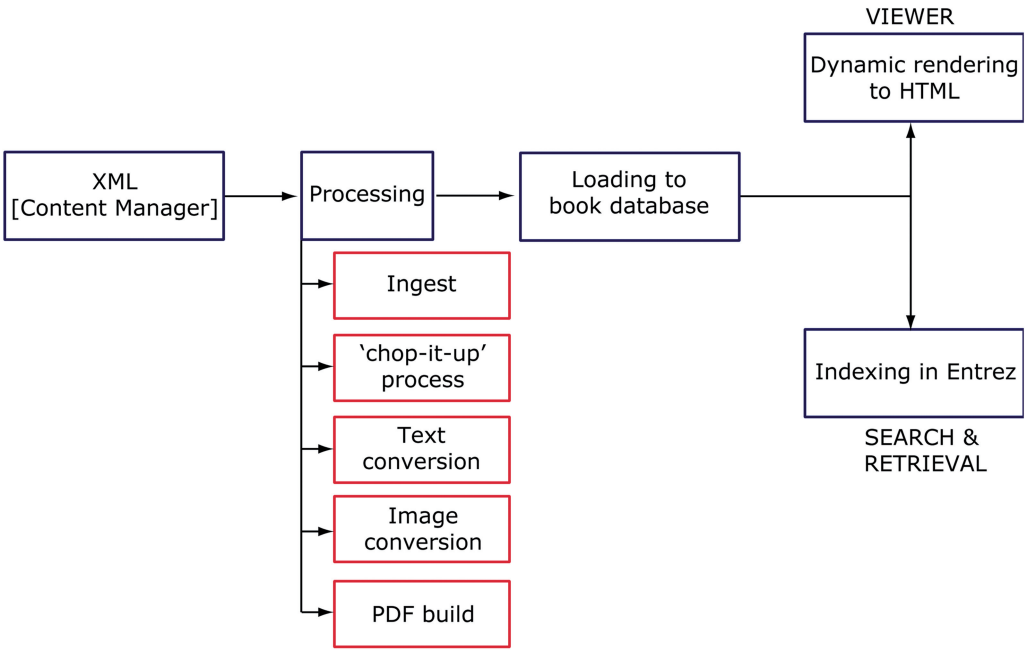


Figure 2. Data processing workflow. XML stored in the Bookshelf CMS goes through several stages of data processing before loading into the book database. Ingest, the first stage of data conversion results in the creation of a tar file package, containing XML, PDF, image and supplementary files. During the 'chop-it-up' process, the single book XML document with root element <book> is chunked into multiple article-like book-part XML documents with root element <book-part>. Text and image conversion processing occurs concurrently. XML files loaded into the book database are dynamically rendered to HTML for viewing in the book viewer application and are indexed in Entrez for search and retrieval.

main database tables. Citations that have PubMed identifiers are stored in the database. The loader creates a unique accession ID, with the 'NBK' prefix for each book-part XML file.

The book database is similar in design to the PMC article database. It is actually a database cluster with a primary database for the main relational tables holding book and book-part information, as well as their properties and attributes, and several secondary blob databases for holding the XML and associated file blobs.

Rendering

Each XML book part is transformed by XSLT to HTML dynamically at run time and delivered to the browser along with the associated files through the book viewer application. The URL for accessing the book part includes the accession ID for that book part.

Performing quality assurance

Quality assurance checks aim to protect the fidelity of data through all stages of processing and ensure accurate rendering and retrieval by the user. Bookshelf uses both manual and automated procedures for performing quality assurance checks.

Indexing in Entrez

Bookshelf records are indexed in Entrez, a text-indexing and data retrieval and discovery application used by major databases at NCBI, such as PubMed and GenBank (6). An indexing script extracts the book-part XML files from the book database and creates an index for each file in an Entrez database. A single book-part XML file may be composed of many units of retrieval (sections, figures, tables) in the indexing database. Each unit is referred to as a DocSum (document summary) having a unique identifier (UID) distinct from the NBK accession ID. Thus, a single book-part accession ID may be represented by many UIDs on retrieval.

There are three types of indexed fields. The first is the key field. Key fields include the Bookshelf NBK accession ID and indexing DocSum UID identifiers, which enable reciprocal linkages between book records and other NCBI databases to be built. References in the book document records (e.g. PubChem identifier) to databases within NCBI are used to build reciprocal links in Entrez between the book document record and the corresponding NCBI database (in this example, PubChem). The reciprocity of linking facilitates information discovery (see Figure 3). The second type of indexed field is the display field. Display fields, e.g. title, display in the document summary for the record during retrieval. Finally, search fields do not display but enable a fielded full-text search. For example, a user can search using the keyword field for all records indexed for a particular keyword or phrase ['pandemic influenza' (keyword)].

SEARCHING, BROWSING AND READING

Books can be searched in Bookshelf or from any NCBI resource page by selecting 'Books' or 'All Databases'. The Entrez application offers a library of search features: use of autocomplete and spell-check dictionaries for queries to improve retrieval, multiple display formats for search results, sort features, query limits, advanced search, clipboard function and log of recent activity. These features are shared with other NCBI resources such as PubMed. Users can search across all books as well as within a specific book (see Figure 4A and B). The Bookshelf browse tool (http://www.ncbi.nlm.nih.gov/books/browse/) is a specialized application, which allows users to filter the list of books by terms or content category selection (see Figure 4C). Book content is delivered to the user in the browser through the book viewer application (see Figure 4D), which features bibliographic and citation information, permits navigation within the book and displays alternate layouts such as print view. Bookshelf content is also accessible through referrals from major search engines.

INTEGRATING BOOKSHELF LITERATURE WITH OTHER NCBI RESOURCES

At NCBI, the literature databases are used to provide annotation for genomic, molecular genetic and structural data resources (7). Book content is being integrated with other NCBI databases such as Gene, Genetic Testing Registry and PubChem to aid in the annotation and discoverability of factual information.

Identifiers to records in other NCBI databases contained within the text of the book (e.g. Gene and PubChem identifiers) when appropriately tagged in the XML data are used to build reciprocal connections between books and the corresponding databases in Entrez (see Figure 3). Source data for books and documents can be highly variable with regard to such tagged references to molecular information; some books and documents contain these references, whereas others do not. The variability is often dependent on the workflow for the particular project: XML-based authoring and publishing workflows in which print and online publications are derived from the same source XML enable rich tagging of references to molecular data, whereas documents that are converted to XML after print publication are often devoid of them. Currently, Bookshelf does not use automated text-mining approaches for capturing references to Gene, Online Mendelian Inheritance in Man (OMIM), GenBank and other molecular databases. Tagging references to molecular data in the XML is only performed by manual or semi-automated processes when there is a high level of confidence that the references created are accurate, and where subsequent quality assurance checks can be performed.

Citation lists in the source XML data are parsed to build links to corresponding citations in PubMed. During the ingest process, the software checks to see whether a PubMed ID has been provided; if not, the software attempts to resolve the citation to a PubMed ID and writes it back into the XML. Citations for select Bookshelf titles are now searchable in PubMed, identifiable by the Books and Documents label in the PubMed

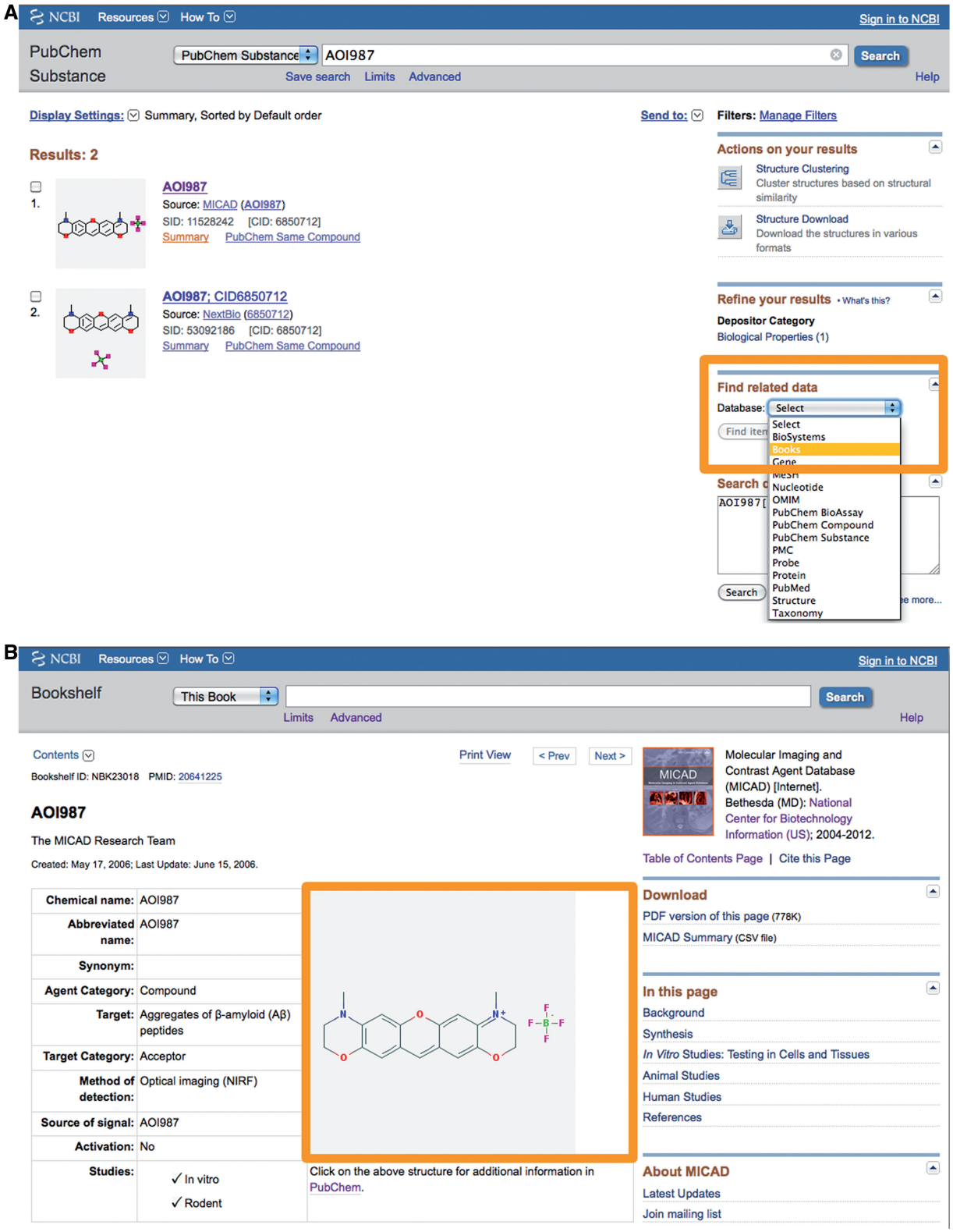


Figure 3. Reciprocal linking between Bookshelf and PubChem. (**A**) A search in PubChem for a substance AOI987 reveals links to book content under 'Find related data'. (**B**) The relevant record in Bookshelf shows the chemical structure for the PubChem substance. This structure renders dynamically in the Bookshelf page from PubChem database. Clicking on the image links to the record in PubChem.

results page. The Related Citations in PubMed feature includes citations to book records. Computation of the similarity relationship for book citations to other articles or books is based on the same algorithm as is used to build similarity relationships between article citations in PubMed (8).

Integration between books and other NCBI resources has also been developed on the basis of curated

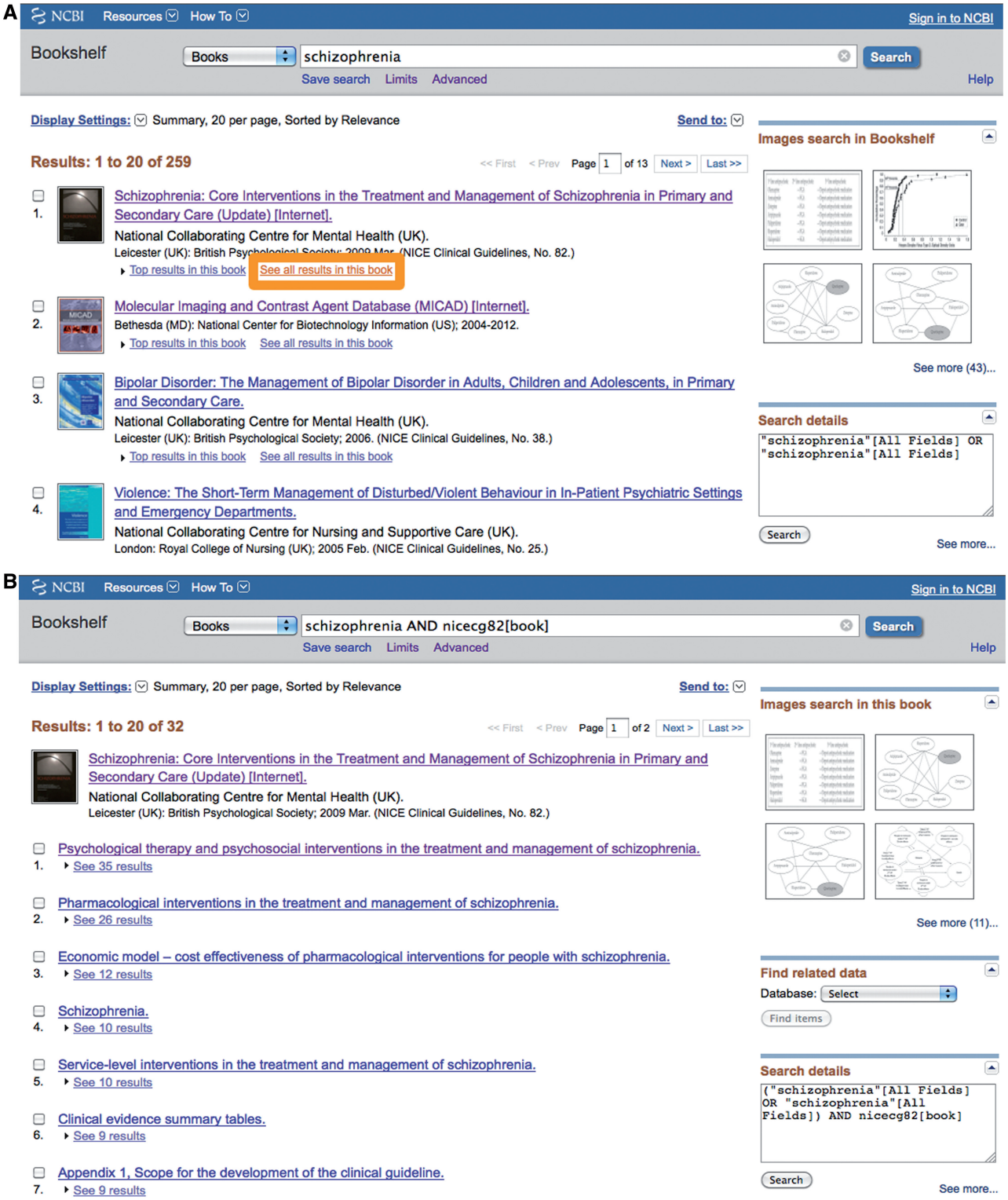


Figure 4. Searching, browsing and reading books. (**A**) Search in Entrez across Bookshelf. Clicking on 'See all results for this book' leads to view shown in B. (**B**) Search within a book. (**C**) Browse tool. Titles can be viewed using term filters or by selection of categories (see highlighted areas). (**D**) Content can be read in the viewer, which displays the citation (orange) and allows for navigation (red) and alternate views (blue).

(continued)

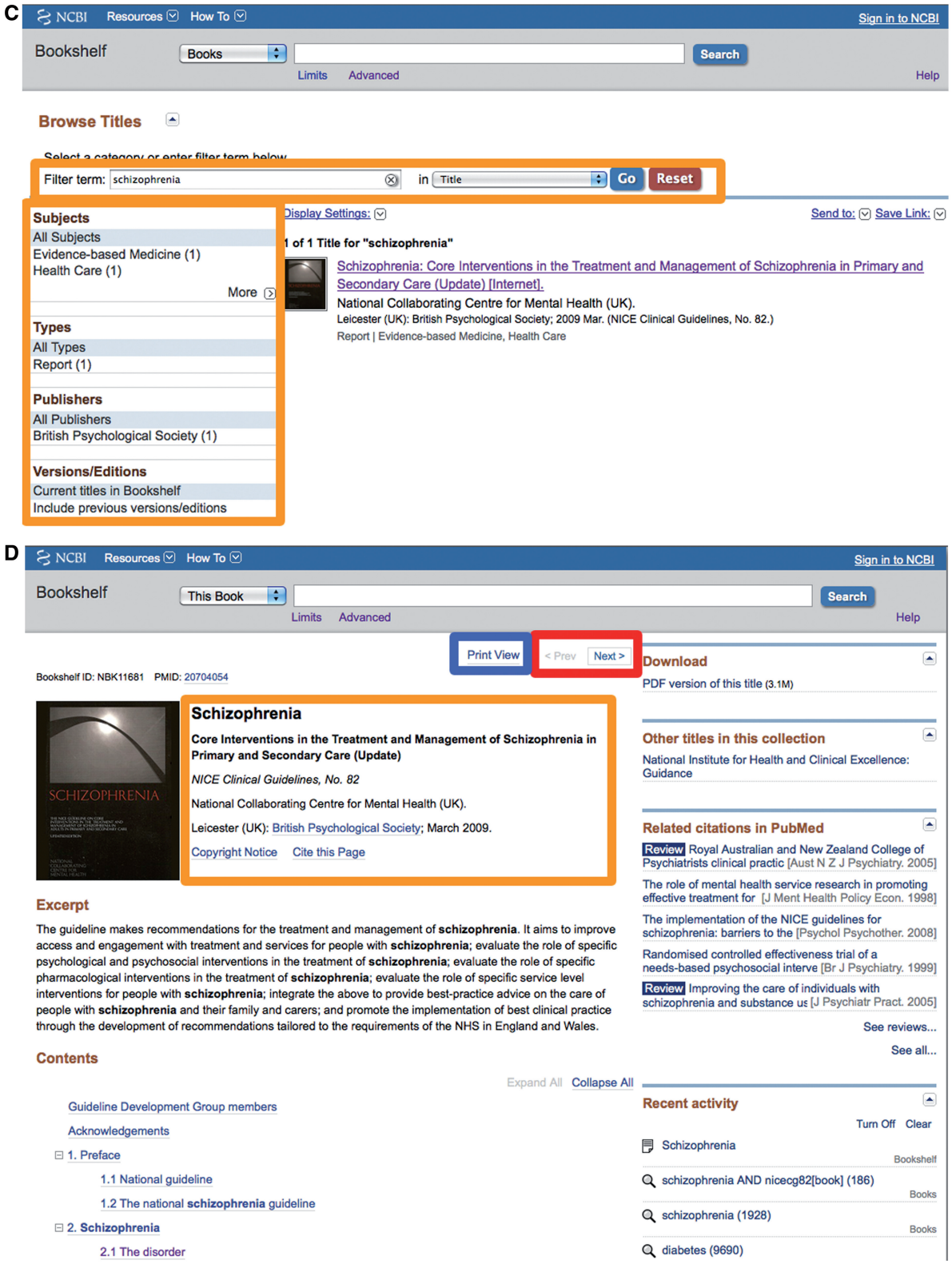
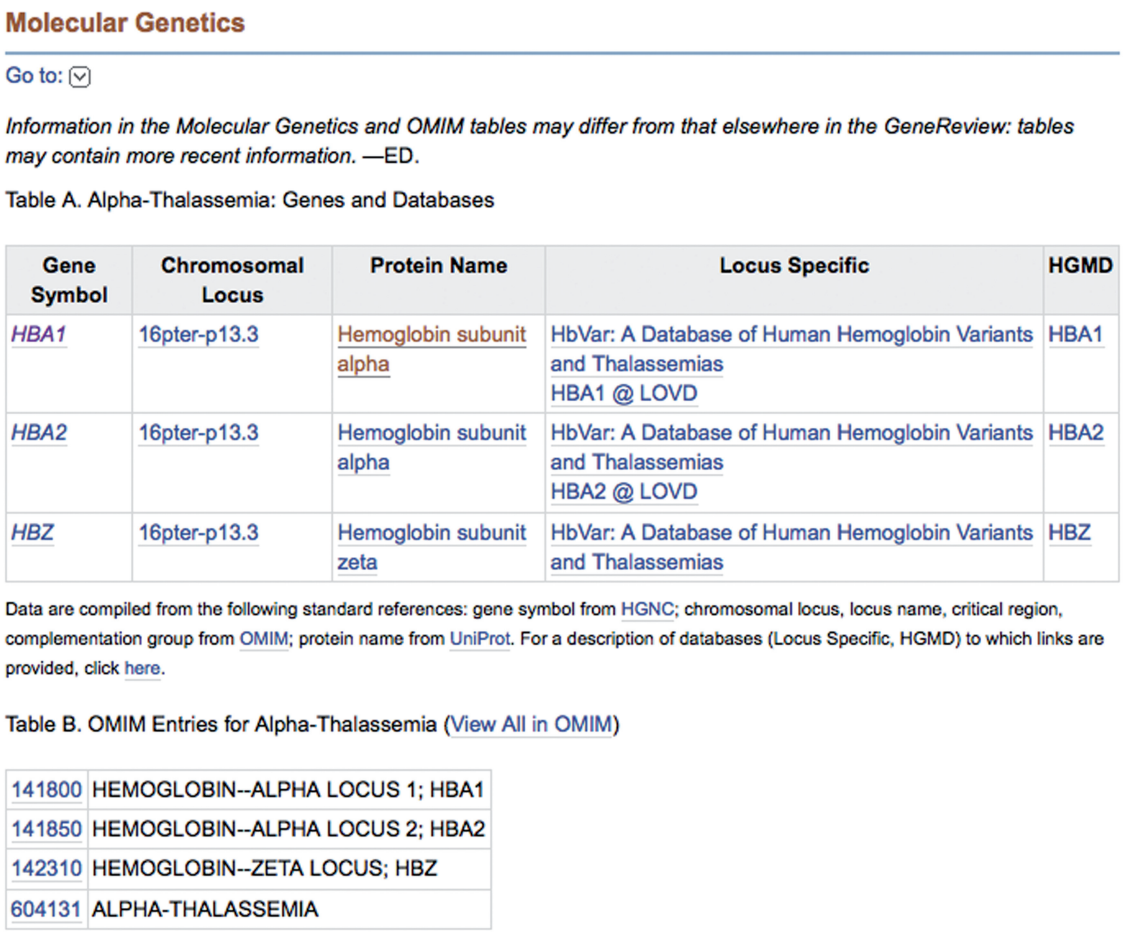


Figure 4. Continued.



Molecular Genetics

Go to:

Information in the Molecular Genetics and OMIM tables may differ from that elsewhere in the GeneReview: tables may contain more recent information. —ED.

Table A. Alpha-Thalassemia: Genes and Databases

Gene Symbol	Chromosomal Locus	Protein Name	Locus Specific	HGMD
HBA1	16pter-p13.3	Hemoglobin subunit alpha	HbVar: A Database of Human Hemoglobin Variants and Thalassemias HBA1 @ LOVD	HBA1
HBA2	16pter-p13.3	Hemoglobin subunit alpha	HbVar: A Database of Human Hemoglobin Variants and Thalassemias HBA2 @ LOVD	HBA2
HBZ	16pter-p13.3	Hemoglobin subunit zeta	HbVar: A Database of Human Hemoglobin Variants and Thalassemias	HBZ

Data are compiled from the following standard references: gene symbol from HGNC; chromosomal locus, locus name, critical region, complementation group from OMIM; protein name from UniProt. For a description of databases (Locus Specific, HGMD) to which links are provided, click here.

Table B. OMIM Entries for Alpha-Thalassemia (View All in OMIM)

141800	HEMOGLOBIN--ALPHA LOCUS 1; HBA1
141850	HEMOGLOBIN--ALPHA LOCUS 2; HBA2
142310	HEMOGLOBIN--ZETA LOCUS; HBZ
604131	ALPHA-THALASSEMIA

Figure 5. Integration of molecular genetic information in books. Data for Table A and B in the Molecular Genetics section of a GeneReviews article are computed based on OMIM entries in a specialized database. The information is dynamically pulled into the text at run time.

relationships built between databases. Even a basic level of expert curation for a book or document may allow for computation of a broader subset of related molecular data, which can be integrated into books, allowing for a larger set of connections between databases to be built. One example elucidating such integration is the GeneReviews (http://www.ncbi.nlm.nih.gov/books/NBK1116/) resource in Bookshelf. Led by Roberta Pagon and her staff at University of Washington, GeneReviews is an expert-authored peer-reviewed collection of reviews on the diagnosis, management and genetic counseling of patients and families with inheritable genetic conditions (9). GeneReviews articles are structured in format and are updated regularly. Relationships between diseases and the genes involved are curated separately in a database designed for the purpose, and updated as that factual information changes, not necessarily when any particular GeneReview is revised. Based on these curated relationships, detailed molecular genetic information from NCBI databases is computationally built nightly for each GeneReviews article. It is stored in the NLM Book DTD format separately until the time of rendering to the browser, when it is integrated into the web page for the GeneReviews article (see Figure 5). This separation of the literature from the molecular genetic data allows the GeneReviews resource to remain as up to date as possible.

NEAR FUTURE

As the collection grows, Bookshelf will need to develop user-centric approaches to information delivery. There will be an increased need to balance comprehensiveness of search results with the usability and relevance of these results. The knowledgebase should also continue to reflect current progress in biomedicine. Bookshelf will be investigating automated text-mining approaches combined with curated and computational approaches with the intent of building molecular data profiles for book records. Such profiles will enable better integration between books and other NCBI databases.

PUBMED CENTRAL INTERNATIONAL AND BOOKS

The PubMed Central International (PMCI) project (http://www.ncbi.nlm.nih.gov/pmc/about/pmci/), a collaboration between NLM, National Institutes of Health, publishers and international agencies that share NLM's goal for archiving the biomedical literature, will soon be broadened to include book content. Currently, there are

two PMCI repositories: EuropePMC [formerly UKPMC (10)] and PMC Canada. The goals for this networked cluster of repositories for archiving and exchanging data are to improve the stability of the archive, facilitate submission of content to local repositories and permit specialized locale-based handling of the biomedical literature, which can be tied to specific health care needs.

ACKNOWLEDGEMENTS

The author thanks Jim Ostell, Ed Sequeira, Laura Dean and Dan Hoeppner for critical reading of the manuscript, comments and constructive discussions; Kathi Canese, Martin Latterner, Vlad Korobtchenko and Sergey Krasnov for data included in Table 1 and Jim Ostell and David Lipman for vision and direction for the Bookshelf project. Several individuals have contributed to the Bookshelf project at NCBI including Laura Dean, Sam Grammer, Stacy Lathrop, Martin Latterner, Ruth Lincoln and Rebecca Orris (Bookshelf team); Jeff Beck, Laura Kelly, Andrei Kolotev, Sergey Koshelkov, Sergey Krasnov, Kathy Kwan, Andrei Lebedev, Chris Maloney, Alexander Maroz, Anh Nguyen, Dima Popov, Vladimir Sarkisov, Ed Sequeira, Karanjit Siyan, Katie Taylor Rose, Kim Tryka and Aleksey Vysokolov (PMC/Literature team); Kathi Canese, Evgeny Kireev, Vlad Korobtchenko, Sharmin Hussain, Aleksey Iskhakov, Vadim Miller, Grisha Starchenko and Abebaw Wubshet (PubMed/NCBI Portal team); Bart Trawick, Greg Schuler, Mark Johnson and Eddie Welker (Public Services); Don Comeau, Zhiyong Lu and John Wilbur (indexing and search); Jennifer Lee and Donna Maglott (Gene); Eric Sayers and Lee Szilagyi (web usage analysis); QA team; Systems team and Dennis Benson, David Gillikin, Diane Boehr and Judith Eannarino. Former NCBIers include Mo Al-Ubaydli, Belinda Beck, Abe Becker, Brooke Dine, Todd Groesbec, Sheila Jiang, Adeline Manohar, Sravanthi Matta, Jo McEntyre and Aleksey Sorokin.

FUNDING

Funding for open access charge: Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Alberts,B., Bray,D., Lewis,J., Raff,M., Roberts,K. and Watson,J.D. (1994) *Molecular Biology of the Cell*, 3rd edn. Garland Science, New York.
2. Dean,L., Orris,R. and Hoeppner,M. (2010) Books with New Looks: The Bookshelf redesign. *NLM Tech. Bull.*, **377**, e9.
3. Dean,L. and Hoeppner,M. (2011) Bookshelf 2011. *NLM Tech. Bull.*, **378**, e8.
4. Beck,J. (2011) NISO Z3996 The Journal Article Tag Suite (JATS): what happened to the NLM DTDs? *J. Electron. Publ.*, **14**, 106.
5. Latterner,M. and Hoeppner,M. (2010) Leafing through XML. In: *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2010*, National Center for Biotechnology Information, Bethesda, MD. http://www.ncbi.nlm.nih.gov/books/NBK47113/.
6. Ostell,J. (2002) Chapter 15: The Entrez Search and Retrieval System. In: McEntyre,J. and Ostell,J. (eds), *The NCBI Handbook [Internet]*, National Center for Biotechnology Information (US), Bethesda, MD. http://www.ncbi.nlm.nih.gov/books/NBK21081/ (10 September 2012, date last accessed).
7. Ostell,J. (2005) Databases of discovery. *ACM Queue*, **3**, 40–48.
8. Lin,J. and Wilbur,W.J. (2007) PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics*, **8**, 423.
9. Pagon,R.A. (2006) GeneTests: an online genetic information resource for health care providers. *J. Med. Libr. Assoc.*, **94**, 343–348.
10. McEntyre,J.R., Ananiadou,S., Andrews,S., Black,W.J., Boulderstone,R., Buttery,P., Chaplin,D., Chevuru,S., Cobley,N., Coleman,L.A. *et al.* (2011) UKPMC: a full text article resource for the life sciences. *Nucleic Acids Res.*, **39**, D58–D65.